DICIONÁRIO ABERTO: UM RECURSO PARA PROCESSAMENTO DE LINGUAGEM NATURAL

Alberto Simões ambs@di.uminho.pt
Rita Farinha rfarinhadp@gmail.com
Universidade do Minho

[Recibido 08-08-09: aceptado 23-11-09]

Resumo

Este artigo apresenta o projecto Dicionário Aberto, a construção de um dicionário aberto, livre e gratuíto, para a língua portuguesa. Para ajudar no arranque optou-se pela transcrição de um dicionário em papel no domínimo público: *Novo Diccionário da Língua Portuguesa, de Cândido de Figueiredo, de 1913*.

Apresentamos o processo usado para a transcrição bem como a metodologia usada para garantir um patamar de qualidade mínima da transcrição, e como o dicionário foi posteriormente convertido para um formato XML, permitindo uma maior facilidade de processamento a terceiros.

Finalmente, são discutidos os problemas existentes no uso de um dicionário com quase um século, e como se pretende proceder à sua modernização (de conteúdos e de grafia), e de que forma este recurso pode ser útil para o processamento da língua portuguesa.

Palabras clave: dicionário aberto, dicionário electrónico, XML. XDXF, transcrição, anotação.

Abstract

This document presents Dicionário Aberto project which aims at the construction of an open-source and free dictionary, for the Portuguese language. To help the bootstrap process, a paper dictionary in the public domain was transcribed: *Novo Diccionário da Língua Portuguesa, of Cândido de Figueiredo, from 1913*.

We will present the transcription process as well as the methodology used to guarantee its quality, and how the document was converted to an XML format, making it easier to other people process the dictionary.

Finally, we will discuss the problems on using a dictionary with almost one hundred years, and how we will proceed with its modernization (of contents and writing), and how this resource can be useful for the Portuguese language processing.

Keywords: dicionário aberto, open dictionary, electronic dictionary, XML, XDXF, transcription, annotation.

1. Introdução

O Dicionário Aberto (DA), disponível na rede em http://www.dicionario-aberto.net/, pretende ser um dicionário electrónico de português moderno, com funcionalidades colaborativas. Para além disso, pretende-se que seja copiável e processável por qualquer pessoa localmente, permitindo assim o seu uso para processamento de linguagem natural.

Dada a quantidade de trabalho necessária para a criação de um dicionário optou-se por proceder à digitalização e transcrição de um dicionário no domínio público. Embora a transcrição também seja um processo demorado e penoso, não são necessários requisitos de formação dos voluntários.

Na secção 2 apresentamos o processo de escolha do dicionário a transcrever, a sintaxe usada para armazenar o documento, e como se criaram ferramentas de validação para se garantir a qualidade da transcrição.

A sintaxe escolhida para a transcrição é demasiado simples, pelo que perde em expressividade. Criou-se uma ferramenta para a conversão e anotação automática deste formato para um subconjunto do XML TEI. Este processo de conversão será discutido na secção 3.

A língua portuguesa tem evoluído e desde a publicação do dicionário usado para a transcrição, foi sujeita a várias reformas, colmatando no recente e polémico acordo ortográfico. Assim, para que se possa tirar partido do dicionário para o processamento da língua recente, estudou-se uma metologia para a modernização da língua. A secção 4 discute os principais desafios e os resultados obtidos num processo automático de actualização da grafia.

Finalmente, a secção 5 apresenta algumas direcções para a exploração deste recurso, desde a extracção de listas temáticas até à extracção de sinónimos, antónimos, ou mesmo de estruturas ontológicas rudimentares.

2. O Projecto Dicionário Aberto

O DA pretende ser um dicionário electrónico, disponível na rede, mas também para uso local, da língua portuguesa (português europeu moderno). A principal justificação para a criação de mais um dicionário da língua portuguesa quando existem outros, como o da Priberam (http://www.priberam.

pt/) ou da Porto Editora (http://www.infopedia.pt/), é a sua vertente aberta e gratuíta. Nenhum dos dicionários citados podem ser usados localmente para consulta ou extracção automática de informação. O uso destes dicionários em qualquer tipo de projectos, desde projectos comerciais a projectos académicos, é completamente impossível.

A construção de um dicionário para uma língua é um processo complicado, demorado e sujeito a erros, além de necessitar de mão de obra especializada. Dado o financiamento nulo deste projecto, a obtenção de mão de obra especializada não foi uma opção.

Decidiu-se usar um dicionário publicado em papel e disponível em domínio público para servir de base ao desenvolvimento do Dicionário Aberto. Resumindo os detalhes da passagem de uma obra ao domínio público[1]: uma obra pode ser considerada no domínio público 70 anos após a morte do autor, ou 70 anos após a edição do documento, o que tiver acontecido mais recentemente.

A escolha recaiu no *Novo Diccionário da Língua Portuguesa*, de *Cândido de Figueiredo*, de 1913. Esta escolha baseou-se na data de morte do autor, data da edição e da disponibilidade de uma cópia digitalizada pela Biblioteca Nacional Digital (http://www.bnd.pt/).

2.1. Processo de Transcrição

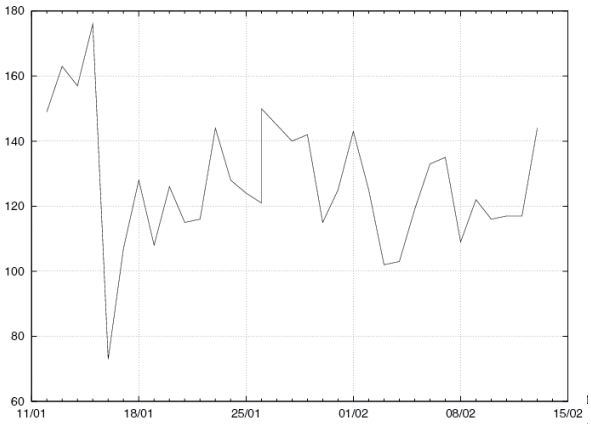


Figura 1: Contribuição diária adicionada ao Dicionário (30 dias no início de 2008).

Para a transcrição deste dicionário usou-se uma aplicação *web* criada para ajudar na transcrição de livros (de qualquer género e de qualquer língua) para integrarem o acervo do *Projecto Gutenberg* (http://www.gutenberg.org/). Esta

[1] Esta é uma descrição simplicista. O processo relativo a direitos de autor é complicado, e deve ser analisado com cuidado. A descrição apresentada, válida na maior parte dos casos, salienta a idade necessária para que um documento seja considerado no domínio público em Portugal.

aplicação está disponível num sítio da Internet, chamado PGDP (*Project Gutenberg, Distributed Proofreaders*, http://www.pgdp.net) (Newby & Franks, 2003).

Este sítio permite que qualquer utilizador possa ajudar na transcrição de um qualquer livro, de um disponível. É apresentada uma página ao utilizador, resultante de um OCR (*Optical Character Recoginition*). O utilizador deve ler e corrigir o texto de acordo com a imagem disponibilizada. É importante salientar a componente de preservação deste projecto, em que se defende que a transcrição realizada seja o mais fiel possível ao documento original.

Esta transcrição é realizada em vários níveis, de forma a que vários utilizadores revejam a mesma página, e deste modo, se possa esperar alguma qualidade na transcrição. No momento de escrita deste artigo encontra em processamento no PGDP a letra **V**.

Os textos, depois de processados no PGDP, são revistos e incorporados no DA. Esta adição de palavras é realizada diariamente, com uma média de 120 palavras diárias. Esta foi uma decisão estratégica, para que permita mostrar aos voluntários que o projecto evolui e o seu trabalho se torna útil. A figura 1 mostra um gráfico relativo à adição de palavras em 30 dias de Janeiro.

A adição diária tem vindo a realizar-se ininterruptamente há cerca de dois anos, sendo que no presente momento se termina a adição da letra **S.**

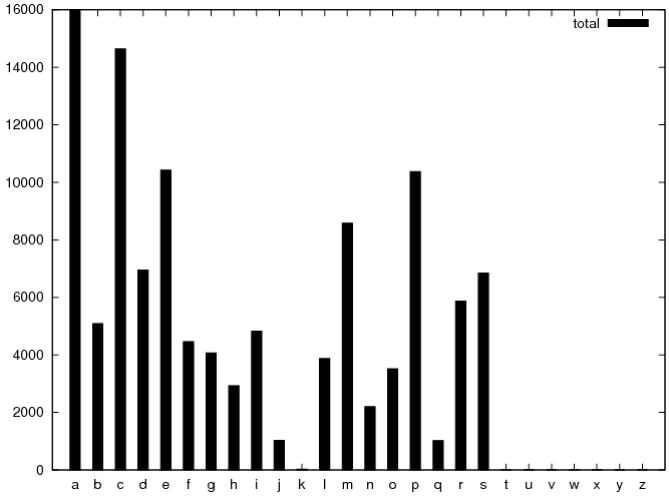


Figura 2: Número de palavras por letra (Agosto de 2009).

A transcrição e incorporação no DA é realizada numa sintaxe puramente textual. Embora esta sintaxe tenha sido definida tendo em conta o formato do dicionário, é demasiado simplista, o que leva a que possam existir algumas ambiguidades no seu processamento automático.

Cachimbo,
_m._
Apparelho de fumador, composto de um fornilho,...
Peça de ferro, em que entra o espigão do leme ...
Buraco, em que se encaixa a vela do castiçal.
* _Bras. de Pernambuco._
Bebida, preparada com aguardente e mel.
* _Pl. Gír._
Pés.
(Do químb. _quixima_)

Sem explicar este formato ao pormenor, salienta-se que cada acepção é colocada numa linha por si só, sendo que algumas são precedidas de uma análise morfológica, de origem geográfica e género de linguagem. Existe também alguma notação de formatação (itálicos e negritos), bem como alguma informação adicional, como o asterisco antes de acepções, que representam acepções novas (que não tinham sido registadas em qualquer outro dicionário).

Optou-se por um formato simplista para permitir que qualquer utilizador do PGDP tivesse possibilidade de colaborar, sem necessidade de aprender uma linguagem de marcação específica, como por exemplo XML.

2.2. Processo de Validação

Dado o cariz colaborativo da transcrição, realizada apenas por voluntários com a mais diversa formação e sensibilidade para a anotação do dicionário, criou-se um conjunto de testes de validação da sintaxe. Estes testes permitiram ao longo dos mais de dois anos de vida do projecto, encontrar vários problemas que foram sendo corrigidos. Estes testes têm sido criados de forma empírica, sempre que se detecta um problema em qualquer entrada. Isto permite que esse tipo de erro possa ser validado em todo o dicionário. Isto leva a que novos erros sejam encontrados e corrigidos.

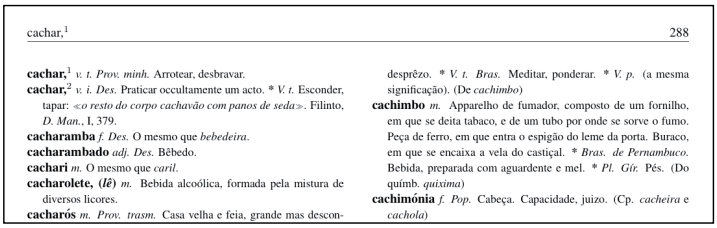
Os testes existentes são, sobretudo, relativos à sintaxe, e incluem desde a simples validação de parêntesis balanceados até ao uso da sintaxe correcta para a definição das análises morfológicas:

- verificar se as marcações de texto em itálico, os parêntesis e as aspas estão balanceadas;
- verificar a existência de espaços duplos no texto, ou de linhas vazias, que atrapalham a divisão automática do ficheiro em verbetes. Este teste também valida a existência de espaços codificados como *non-breaking-spaces* em unicode, que devem ser substituídos por espaços convencionais;
- verificar que as reticências dos exemplos de uso estão entre as marcas de texto itálico (este teste é especialmente importante uma vez que as regras gerais de transcrição no PGDP sugerem que as reticências não sejam in-

cluídas nas marcas de texto itálico quando no fim ou início dessa zona de texto);

- assegurar que as ligações entre verbetes (construções como *ver também* ou *comparar com*) incluem apenas os termos relacionados, e não pontuação, para que seja automatizavel a criação de hiper-ligações;
- dada a impossibilidade dos voluntários introduzirem todo o tipo de caracteres, alguns são codificados manualmente (por exemplo, um ẽ é codificado como **[~e]**). Logo, é importante verificar se todos estes caracteres foram convertidos para a respectiva codificação unicode;
- confirmar que não existem acepções quebradas (ou seja, que não existem linhas que iniciem com letra minúscula);
- verificar a sintaxe da primeira linha do verbete, que deve ser constituída por: um conjunto de caracteres opcionais, classificadores do tipo de entrada; o termo em causa; o número de acepção (opcional); e informação fonética (também opcional).

Sempre que um destes testes falha é criado um relatório indicado o ficheiro e linha que deve ser corrigido, e qual o problema encontrado. Este relatório é produzido de forma a que no caso de se encontrarem muitos testes se possa proceder a uma correcção automática.



cachar,[1] 288

cachar,[1] *v. t. Prov. minh.* Arrotear, desbravar.
cachar,[2] *v. i. Des.* Praticar occultamente um acto. * *V. t.* Esconder, tapar: «*o resto do corpo cachavão com panos de seda*». Filinto, *D. Man.*, I, 379.
cacharamba *f. Des.* O mesmo que *bebedeira*.
cacharambado *adj. Des.* Bêbedo.
cachari *m.* O mesmo que *caril*.
cacharolete, (*lê***)** *m.* Bebida alcoólica, formada pela mistura de diversos licores.
cacharós *m. Prov. trasm.* Casa velha e feia, grande mas desconfortável. * *V. t. Bras.* Meditar, ponderar. * *V. p.* (a mesma significação). (De *cachimbo*)
cachimbo *m.* Apparelho de fumador, composto de um fornilho, em que se deita tabaco, e de um tubo por onde se sorve o fumo. Peça de ferro, em que entra o espigão do leme da porta. Buraco, em que se encaixa a vela do castiçal. * *Bras. de Pernambuco.* Bebida, preparada com aguardente e mel. * *Pl. Gír.* Pés. (Do químb. *quixima*)
cachimónia *f. Pop.* Cabeça. Capacidade, juizo. (Cp. *cacheira* e *cachola*)

Figura 3: Extracto de uma página do DA em formato PDF.

2.3. Disponibilização do DA

Como motivação para o processo de transcrição decidiu-se, como já foi referido, adicionar um conjunto de palavras diariamente, com o intuito de mostrar crescimento e utilidade.

Procedeu-se à disponibilização do DA na rede, de diferentes formas:

- **formato texto:** uma das principais críticas aos restantes dicionários disponíveis na rede é o facto de não serem processáveis localmente. Um dos formatos disponibilizados desde a criação do projecto é o documento textual na sintaxe original (descrita anteriormente). Como será discutido na secção 3 também se encontra disponível o dicionário em formato textual etiquetado em TEI.
- **formato PDF:** embora se não sugira a impressão do dicionário como um todo em papel por razões económicas e ecológicas, é disponibilizado um documento PDF em formato dicionário (com cabeçalhos típicos de

impressão, e duas colunas), de acordo com a figura 3. Este formato não é mais do que uma demonstração da viabilidade de se editar em papel o DA em caso de interesse/necessidade.

- **pesquisa na rede:** o uso típico de um dicionário na rede é através da sua pesquisa directa, introduzindo uma palavra ou termo, e obtendo todas as acepções relacionadas. Além desta funcionalidade típica, no DA é apresentado um conjunto de termos com o mesmo prefixo, e também um conjunto de palavras semelhantes com uma distância de *Levenshtein* de 1 (Levenshtein, 1966).
- **navegação na rede:** é nossa convicção que a pesquisa num dicionário convencional não se reduz à consulta do verbete relativo à palavra em causa, mas que o carácter curioso do ser humano leva a que este acabe por ler três ou quatro verbetes na vizinhança da palavra pesquisada. Isto levou a que se implementasse uma ferramenta de pesquisa alternativa que, para além de apresentar o verbete procurado, apresenta uma lista de 15 entradas vizinhas que ocorrem antes da palavra em causa, bem como as 15 entradas vizinhas que ocorrem depois da palavra.
- **dicionário StarDict:** embora cada vez mais se defenda o uso das aplicações na rede, continuam a existir muitas ocasiões em que se pretende consultar um dicionário e não temos acesso à Internet. Nestes casos é imprescindível ter acesso a um dicionário local. Embora se pudesse usar o formato textual do dicionário ou a sua versão PDF, o seu uso é ineficiente e pouco atractivo. O DA providencia uma versão do dicionário para uso local através da ferramenta StarDict (ver exemplo na figura 4).

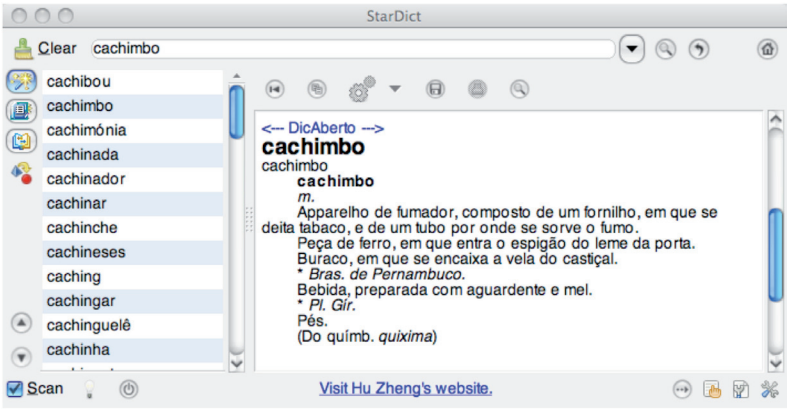


Figura 4: Exemplo de execução do StarDict.

3. Anotação do DA em XML.

O formato definido para a transcrição do DA é simples de entender e de usar para a transcrição, mas não é versátil nem tem expressividade suficiente

para se armazenar o dicionário, e para que possa ser processado de forma automática, já que não há marcação explícita de certos elementos como origem etimológica, análise morfológica, origem geográfica, etc.

Para colmatar este problema decidiu-se converter o DA num formato baseado em XML. Esta secção está dividida em duas partes. Na primeira iremos discutir alguns formatos XML disponíveis para a anotação de dicionários, e usados em diferentes tipos de projectos, terminando com uma proposta de um subconjunto para o caso específico do DA. A segunda parte apresenta o processo automático utilizado na conversão do formato textual para XML, bem como os testes desenvolvidos que nos permitem assegurar alguma qualidade na anotação realizada.

3.1. Definição do sub-formato TEI

Existem algumas iniciativas de definição de formatos XML para dicionários, como o XDXF (XML Dictionary Interchange Format), mas que não são suportadas por nenhuma instituição ou organização, e que (no momento de escrita deste documento) demonstram falta de organização e profissionalismo na documentação disponibilizada.

Suportado por um consórcio, existe o meta-formato TEI (Vanhoutte, 2004) (*Text Encoding Initiative*). Este meta-formato foi definido para a codificação em XML de qualquer tipo de documento, desde livros a corpora, passando por dicionários e glossários. Chamam-lhe meta-formato já que é composto por uma base comum partilhada por diferentes módulos que podem ser usados independentemente para formatar diferentes tipos de documentos, obtendo desta forma modularidade.

Existem outros projectos que têm vindo a construir dicionários e utilizando o XML para os formatar. Um exemplo é o *Dicionário de Dicionários da Língua Galega* (Santamarina, 2003). Dado que este projecto pretende armazenar informação de dicionários de diferentes anos e com diferente tipo de informação, o formato XML definido é razoavelmente complicado, e a sua adopção iria gerar alguma confusão.

O *Dicionário Inglês-Galego* (Álvarez Lugrís, 2008; Gómez Guinovart, Díaz Rodriguez e Álvarez Lugrís, 2008) também usa um formato XML, mas bastante mais simples do que o do TEI ou do Dicionário de Dicionários, e que pode ser visto como um subconjunto do TEI, com alguma reorganização pela sua natureza bilingue.

O formato que usamos é um subconjunto do TEI, tendo sido removida alguma complexidade do TEI oficial (o que torna os ficheiros XML incompatível com o DTD do TEI, sendo que a sua conversão para o formato oficial não é complicada).

O formato usado divide o dicionário em verbetes que são etiquetados com a etiqueta *entry*. Dentro desta etiqueta aparecem habitualmente três zonas diferentes:

- *form*, apresenta as diferentes formas do termo em causa, nomeadamente a forma escrita e a sua leitura fonética (completa ou parcial);
- *sense*, delimita as várias acepções. A detecção automática de acepções é praticamente impossível, pelo que se optou por distinguir apenas aquelas em que a informação de Part-Of-Speech difere. Além destas, também se distinguiram aquelas que por algum motivo foram anotadas pelo autor do dicionário, e portanto, de fácil extracção.
- *etym*, apresenta a origem etimológica da palavra.

A principal diferença desta estrutura com a usada pelo TEI é que se optou por não subdividir algumas destas secções como seria sugerido. Dentro de cada acepção (etiqueta *sense*), delimitou-se apenas a informação morfológica, a origem geográfica do termo, o domínio a que pertence, bem como o género de linguagem. Embora esta informação tenha sido etiquetada de uma forma simplicista durante a transcrição do dicionário, um conjunto de heurísticas e algum trabalho manual permitiram fazer a sua anotação.

3.2. Processo de Anotação e Teste

O processo de anotação em XML foi baseado num sistema de reescrita (Almeida e Simões, 2001). Este sistema permite a reescrita de documentos textuais de uma forma simples, legível e eficiente. O sistema processa um conjunto de regras compostas por padrões, e acções a tomar no caso de o padrão estar presente no texto.

```
padrão ==> acção   !! condição sobre o padrão
```

Este sistema permitiu a escrita de padrões simples sobre o documento a processar, que vão retirando informação do texto. Por exemplo, é simples de detectar os termos de cada verbete uma vez que se encontram entre asteriscos. Depois de detectar o termo, será simples de reconhecer a zona de transcrição fonética. Este processo itera sobre o texto original, extraindo progressivamente a informação mais simples de detectar, até à informação mais complexa.

Para alguns casos específicos foi necessária a construcção de listas de palavras especiais. Por exemplo, para detectar o domínio a que pertence determinado termo foi necessário criar uma lista das abreviaturas usadas pelo autor para representar esses domínios. Embora esta lista exista no início do dicionário (juntamente com outras abreviaturas usadas), foram detectadas novas abreviaturas que, dada a construção manual do dicionário, acabaram por não serem colocadas na tabela de abreviaturas.

Segue-se um pequeno exemplo da entrada da palavra *cachimbo* em XML TEI.

```
<entry>
<form><orth>Cachimbo</orth></form>
<sense>
```

<gramGrp>m.</gramGrp>
<def>
Apparelho de fumador, composto de um fornilho, em que se deita tabaco, e de um tubo por onde se sorve o fumo.
Peça de ferro, em que entra o espigão do leme da porta.
Buraco, em que se encaixa a vela do castiçal.
</def>
</sense>
<sense ast="1">
<usg type="geo">Bras. de Pernambuco.</usg>
<def>Bebida, preparada com aguardente e mel.</def>
</sense>
<sense ast="1">
<gramGrp>Pl.</gramGrp>
<usg type="lang"> Gír </usg>
<def>Pés.</def>
</sense>
<etym ori="químb">(Do químb. _quixima_)</etym>
</entry>

Temos consciência de que esta anotação não é perfeita. No entanto, e dada a abordagem automática que foi aplicada, parece-nos suficiente para a maior parte das necessidades de investigação. Por outro lado, é simples a adição de novas regras ao sistema de reescrita, de modo a permitir o tratamento da notação em falta.

Para testar a conversão optou-se por validar apenas o formato XML: verificar se a primeira linha de cada documento é a definição típica dos documentos XML que indicam a codificação usada, e verificar o correcto aninhamento das etiquetas XML (ou seja, a validação típica de que o documento esteja bem formado). Também é realizada uma validação superficial à informação gramatical de todas as entradas, garantindo que estão de acordo com a notação escolhida.

Este processo de conversão para XML e posterior validação levou a que vários pequenos problemas fossem detectados e corrigidos nos ficheiros transcritos, e que de outro modo seriam difíceis de detectar.

4. Actualização da grafia do Português

O processo de modernização ou actualização da grafia do dicionário ainda não foi aplicado a todo o dicionário, uma vez que o dicionário ainda não foi todo transcrito. No entanto, tem-se vindo a realizar experiências, criação de regras e análise de cobertura.

Antes de realizar a actualização da grafia é necessário definir quais as áreas do dicionário que devem ser modernizados:

- dada a existência de origens etimológicas é natural que apareçam palavras noutras línguas, como latim, hebraico, grego, inglês, fran-

cês e outros. Estas palavras não devem ser alteradas pelo sistema automático de modernização;

- do mesmo modo, é necessário distinguir o que são palavras do que são abreviaturas. Enquanto que num texto obtido de um artigo jornalístico ou de um romance as abreviaturas são mais ou menos previsíveis e detectáveis (dada a sua pouca afluência), num dicionário essa tarefa torna-se bastante mais complicada dada a profusão de diferentes abreviaturas;
- os exemplos de uso, citações de livros, etc, estão com a grafia original (não necessariamente a mesma usada no dicionário), o que leva a que se torne mais complicada a sua modernização: algumas das regras gerais que podemos aplicar ao dicionário não podem ser aplicadas a estes exemplos;
- algumas palavras deixaram de existir, e não existe uma versão *moderna*. Simplesmente, morreram. Nestes casos o sistema não irá conseguir modernizar a palavra, e irá mantê-la na sua forma original. Será preciso um processo de detecção destas palavras para a sua devida anotação.

Depois de resolver estes problemas torna-se possível aplicar heurísticas de modernização. Algumas regras são relativamente simples, como a remoção de um **l** em palavras como "cavallo", ou a transformação do par **ph** na letra **f**. Embora estas transformações sejam certeiras, é preciso decidir o que fazer com os verbetes resultantes, uma vez que algumas palavras já existem no dicionário com ambas as grafias e com definições iguais ou ligeiramente diferentes.

Outras transformações, como a remoção de acentos de palavras como "mulhér" ou "fórma" pode ser automatizada com base num dicionário de português moderno. Se, após a remoção do acento, a nova palavra existe num dicionário actual, então é provável que essa transformação esteja correcta.

Neste momento está-se a proceder a estudos de ocorrência de palavras, criando histogramas com contagens. Deste modo serão detectadas as palavras mais usadas e, portanto, que mais contribuem para a leitura do dicionário.

A tabela 1 apresenta algumas das substituições que têm vindo a ser escritas e testadas[2]. Junto a cada uma é apresentado o número de palavras constantes no dicionário (não apenas termos). Importante também referir que estas substituições só são feitas se a palavra não existir no dicionário[3].

[2] Nesta tabela a barra vertical representa o fim ou o início da palavra.

[3] No caso concreto desta experiência foi usado um dicionário de um corrector ortográfico, com cerca de 43 mil lemas, o que não garante a existência de palavras menos comuns como *coleóptero* ou *liliácea*.

ph	f	544	ó	o	261	ll	l	1119	ô	o	422
y	i	550	ê	e	329	mm	m	434	é	e	197
cc	c	321	pç	ç	14	bb	b	10	nn	n	190
pp	p	293	aes|	ais	226	ff	f	309	ê	é	126
ý	í	38	Ch	qu	174	tt	t	151	ò	o	36
ehe	ee	73	|sc	c	44	sç	ç	56	í	i	471
mpt	nt	53	an|	ã	43	ct	t	205	ahi	aí	30

Tabela 1: Substituições mais comuns e número de palavras afectadas.

As substituições até agora definidas recuperam cerca de 9623 palavras em grafia antiga para a correspondente grafia moderna, de entre 146079 formas. Considerando o número total de palavras que ocorrem no dicionário (mais de um milhão), 18% eram desconhecidas à partida, tendo este valor descido para 10.5% (ou seja, 88.5% das palavras seriam convertidas para português actual).

Este processo ainda não está terminado, havendo ainda um conjunto de substituições possíveis que ainda não foram analisadas, bem como um conjunto de abreviaturas que ainda não foram reconhecidas. Existem também erros de transcrição que não são ainda detectados de forma automática.

5. Aplicações do DA no Processamento da Língua Portuguesa

A existência de um dicionário livre da língua portuguesa permite que se possa processar localmente para a extracção de relações entre palavras. Por exemplo, é simples a construção de listas de animais (quadrúpedes, aves, peixes), listas de plantas, etc, bastando para isso a pesquisa de determinadas expressões na definição. Se a definição começar por "*ave*" ou "*pássaro*" podemos considerar que a entrada se refere a uma ave.

Do mesmo modo podem-se extrair sinónimos encontrando verbetes cuja definição inclua o padrão "*o mesmo que ...*"

Usando estas heurísticas simples construiram-se automaticamente listas com 692 aves, 435 peixes, 2101 plantas, ou 13856 conjuntos de sinónimos.

A generalização destes padrões e uma análise mais detalhada das construções típicas usadas pelo dicionarística permitirão a extracção de relações mais ricas.

6. Conclusões e Trabalho Futuro

É nossa convicção que este trabalho é útil e relevante em vários contextos. Por um lado a análise lexicográfica contrastiva (como se escrevia e como se escreve), a análise dicionarística (qual a evolução das definições de palavras durante o tempo), a análise literária (já que muitos livros, por exemplo

de Camilo, usam palavras não constantes em dicionários actuais) e a análise e processamento automático para extracção de conhecimento (ontologias, definições, etc).

Pretendemos que o projecto mantenha a transcrição original acessível a todos os utilizadores, uma vez que é património. No entanto, será desenvolvido um sistema *Wiki* sobre este dicionário que permita a qualquer utilizador adicionar palavras, adicionar acepções, e alterar definições. Para garantir que uma qualidade mínima é mantida, este *Wiki* será baseado numa cascata de permissões, em que qualquer definição ou palavra só se torna oficial quando validada por um utilizador responsável.

A transformação quer da versão original, quer da versão moderna, num formato aberto como o TEI permite que qualquer investigador em PLN possa processar o dicionário sem qualquer limitação.

Embora a modernização da língua portuguesa possa parecer neste momento um sonho vago, parece-nos que será possível modernizar 20% das entradas (das palavras mais comuns) de forma automática. Note-se que neste momento (letra P) o dicionário está a rondar as 100.000 entradas, pelo que 20% corresponde a 20.000 entradas.

Finalmente, e em termos de finalização de transcrição, o processo encontra-se neste momento bem oleado e há previsões de que a transcrição do dicionário termine este ano civil. É provável que a revisão final venha a demorar mais uns meses, antes de ser publicado na Internet.

REFERÊNCIAS BIBLIOGRAFICAS

Almeida, J.J. & Simões, A. 2001. *Text-to-speech: A rewriting system approach*. Procesamiento del Lenguaje Natural, 27: 247–253.

Álvarez Lugrís, A. 2008. *O dicionario CLUVI inglés-galego*. Longa Lingua, 20.

Guinovart, X.G. & Rodríguez, E. D. & Lugrís, A.A. 2008. *Aplicacións da lexico-grafía bilingüe baseada en córpora na elaboración do dicionario CLUVI inglés-galego.* Viceversa: Revista Galega de Traducción, 14.

Levenshtein, V.I. 1966. *Binary codes capable of correcting deletions, insertions, and reversals*. Soviet Physics Doklady, 10:707–710.

Newby, G.B. & Franks C.C. 2003. *Distributed proofreading*. International Conference on Digital Libraries, páginas 361–363.

Santamarina, A. 2003. *Diccionario de diccionarios*. CD-ROM. A Coruña: Fundación Pedro Barrié de la Maza / Instituto da Lingua Galega.

Vanhoutte, E. 2004. *An introduction to the TEI and the TEI consortium*. Lit Linguist Computing, 19(1):9–16, April.